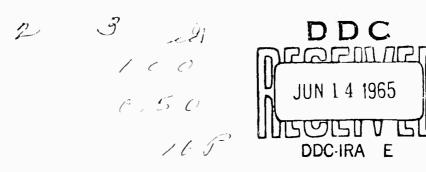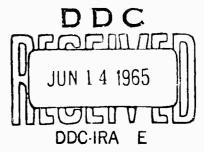NOTE ON A NEW COMPUTATIONAL DATA SMOOTHING PROCEDURE
SUGGESTED BY MINIMUM MEAN SQUARE ERROR ESTIMATION

P. Swerling

May, 1965

NOTE ON A NEW COMPUTATIONAL DATA SMOOTHING PROCEDURE
SUGGESTED BY MINIMUM MEAN SQUARE ERROR ESTIMATION

P. Swerling[*]

Consultant to The RAND Corporation, Santa Monica, California

I. INTRODUCTION

Reference 1 contains a discussion of some possible computational procedures for finding the exact minimum mean square error (MMSE) estimates of parameters in the case where the observations depend non-linearly on the unknown parameters. The procedures analyzed there are for the case where the observation noise is Gaussian and additive.

It was pointed out in that paper that the applicability of such estimation procedures, that is, those that would yield MMSE estimates if the noise were Gaussian and additive, is not necessarily limited to cases where the noise is actually Gaussian or additive. An analogy was made to least squares procedures, which would be maximum likelihood if the noise were additive and Gaussian, but whose applicability is not restricted to such cases (although if these conditions are not satisfied one can no longer necessarily make the same claims regarding optimality).

In contrast to least squares procedures, the MMSE procedures achieve exact (not merely asymptotic) minimum mean square estimation error even when the observational data depend non-linearly on the unknown parameters.

Also, the computational procedures are quite different. In the least squares procedures, not only is the optimality criterion the minimization of the statistically expected square of the error, but the computational procedure involves minimizing, with respect to the parameters, a function of both the observed data and the parameters; and this function is a weighted sum of squares, or more generally a positive quadratic form, in the residuals.

In contrast, while the optimality criterion for the MMSE procedures is still the minimization of the statistically expected square error, the computational procedures do not necessarily involve the minimization of a function of observed data and parameters; rather, they involve the evaluation of certain integrals over the parameter space. Although the computational implementation of MMSE estimation is often very difficult, there may be cases in which the evaluation of the necessary integrals is more convenient, even from the computational point of view, than the minimization required in least squares procedures.

In the "strong signal" case the least squares procedure can be linearized; that is, the functional dependence of the observed data on the unknown parameters can be treated as linear within some region surrounding the true parameter values, and other regions of the para-

meter space can be effectively ignored. In such cases, the estimates obtained are asymptotically the same as those obtained from MMSE procedures.

More generally, if the logarithm of the likelihood function is convex in the unknown parameters in a region surrounding the true values, for sufficiently strong signal everything other than this region may be ignored, and iterative techniques may be applied to find the least squares estimate.

The purpose of this note (following a suggestion made in Ref. 1) is to investigate the form taken by MMSE estimation procedures in the case where the dependence of the observed data on the parameters can, with good approximation, be considered to contain only constant, linear, and quadratic terms. The resulting procedures can be considered as alternatives to the iterative procedures used in least squares.

II. MMSE ESTIMATES WITH ADDITIVE GAUSSIAN NOISE

It is assumed that the observed data can be represented in the form

$$S(t) = f(x_1, \ldots, x_n, t) + \mathcal{E}(t) \tag{1}$$

where t is a time parameter (actually, it need only be assumed that t is a parameter whose value is known--it need not be interpreted as time, but it is convenient to speak of it in this way); $x_1, \ldots, x_n$ are n real, unknown parameters; and $\{\mathcal{E}(t)\}$ is the observation error.

We will write

$$x = (x_1, \ldots, x_n) \tag{2}$$
$$x \in X$$

where X is a parameter set, and n is the number of unknown real parameters.

The parameter vector x is assumed to have an a-priori probability density function $p(x)$ over X.

Also, it will be assumed initially for purposes of derivation that the observed data consist of samples of $S(t)$ at a number N of discrete time points $\{t_\mu\}$, $\mu = 1, \ldots, N$, so that the observed data are

$$S_\mu = S(t_\mu) = f(x, t_\mu) + \mathcal{E}(t_\mu), \mu = 1, \ldots, N \tag{3}$$

Eventually, however, the results will also be obtained, by a limiting process, for cases where the function $S(t)$ is continuously observed over some interval.

Suppose the set of random variables $\{\mathcal{E}_\mu\}$ are Gaussian, have zero means, and have covariance matrix

$$\overline{\mathcal{E}_\mu \mathcal{E}_\nu} = \varphi_{\mu\nu} = \varphi(t_\mu, t_\nu) \tag{4}$$

where

$$\varphi(s, t) = \overline{\mathcal{E}(s)\,\mathcal{E}(t)} \tag{5}$$

Then, the joint probability density of $\{S_\mu\}$, conditional on the parameters having given values, is

$$p(S \mid x) = (2\pi)^{-\frac{N}{2}} \Delta^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2} \sum_{\mu,\nu=1}^{N} \eta_{\mu\nu} [S_\mu - f(x, t_\mu)][S_\nu - f(x, t_\nu)]\right\} \tag{6}$$

where

$$\begin{aligned}
S &= (S_1, \ldots, S_N) \\
x &= (x_1, \ldots, x_n) \\
\Delta &= \det |\varphi| \\
\eta_{\mu\nu} &= (\varphi^{-1})_{\mu\nu}
\end{aligned} \tag{7}$$

If the a-priori probability density of x over X is $p(x)$, then the joint p.d.f. of S and x is

$$p(S, x) = p(S \mid x)\, p(x) \tag{8}$$

with $p(S \mid x)$ given by Eq. (6). (These various density functions do not necessarily have the same functional form, even though the symbol p is used in each case; ambiguity is removed by noting the argument of p.)

Let $\beta(x) = \beta(x_1, \ldots, x_n)$ be any real-valued function of $x_1, \ldots, x_n$ having finite second moment with respect to $p(x)$.

Then, the MMSE estimate of $p(x)$, denoted by $\hat{\beta}$, is that function of S which minimizes

$$\int_{\Omega} \int_{X} [F(S) - \beta(x)]^2 \, p(S, x) \, dS \, dx$$

for all functions $F(S)$ of S. Here, Ω is the set consisting of the possible values of the vector S.

It is then easy to show that the MMSE estimate $\hat{\beta}$ is[1]

$$\hat{\beta} = \frac{\int_{X} \beta(x) \, p(x) \, \exp\left\{-\tfrac{1}{2} A(x) + u(x)\right\} \, dx}{\int_{X} p(x) \, \exp\left\{-\tfrac{1}{2} A(x) + u(x)\right\} \, dx} \tag{9}$$

where

$$A(x) = \sum_{\mu,\nu=1}^{N} \eta_{\mu\nu} \, f(x, t_\mu) \, f(x, t_\nu) \tag{10}$$

$$u(x) = \sum_{\mu,\nu=1}^{N} \eta_{\mu\nu} \, f(x, t_\mu) \, S(t_\nu) \tag{11}$$

If the observational data consists of the function $S(t)$ observed continuously over an interval (T_1, T_2), then $\hat{\beta}$ is given by Eq. (9) but with $A(x)$ and $u(x)$ in Eqs. (10) and (11) replaced by their limiting values as $N \to \infty$ and as the points $\{t_\mu\}$ become dense in (T_1, T_2).

Methods of evaluating such limits are discussed at length in Refs. (2) - (4). It might be noted that $A(x)$ is essentially a kind of signal-to-noise ratio appropriate to noise having the covariance function $\varphi(s, t)$, while $u(x)$ is a linear operation on the received signal which is equivalent to that generalization of cross-correlation which is appropriate to noise having the covariance function $\varphi(s, t)$.

In fact, in the white-noise case, in the limit

$$A(x) = \frac{2}{N_o} \int_{T_1}^{T_2} f^2(x, t) \, dt \tag{12}$$

$$u(x) = \frac{2}{N_o} \int_{T_1}^{T_2} f(x, t) \, S(t) \, dt \tag{13}$$

where N_o = one-sided noise spectral density.

Now, we will consider the expansion of both $u(x)$ and $A(x)$ through second order in $\{x_i\}$. In essence, we will expand $f(x, t)$ through second order in $\{x_i\}$ for purposes of approximating $u(x)$. However, in approximating $A(x)$, the full expansion of $f(x, t)$ to second order will not be used; rather, $A(x)$ will itself be expanded through at most second order.

This can be expected to yield very accurate approximations to $A(x)$ in many cases of interest, since there are many such cases in which $A(x)$ is either independent of x or depends on x in a way which can be well approximated by retaining only terms through quadratic.

Thus, suppose $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ is some definite value of x and that in the neighborhood of $\bar{x}$

$$f(x, t) = f(\bar{x}, t) + \sum_{i=1}^{n} \frac{\partial f(\bar{x}, t)}{\partial x_i} (x_i - \bar{x}_i) \tag{14}$$

$$+ \tfrac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2 f(\bar{x}, t)}{\partial x_i \, \partial x_j} (x_i - \bar{x}_i)(x_j - \bar{x}_j)$$

$$A(x) = A(\overline{x}) + \sum_{i=1}^{n} \frac{\partial A(\overline{x})}{\partial x_i} (x_i - \overline{x}_i) \tag{15}$$

$$+ \tfrac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2 A(\overline{x})}{\partial x_i \, \partial x_i} (x_i - \overline{x}_i)(x_j - \overline{x}_j)$$

It may be noted that the derivatives of $A(x)$ can easily be evaluated in terms of those of $f(x, t)$ via Eq. (12), in the white noise case, or more generally via Eq. (10).

For notational convenience, let

$$\frac{\partial f(\overline{x}, \, t)}{\partial x_i} = d_i(\overline{x}, \, t) \tag{16}$$

$$\frac{\partial^2 f(\overline{x}, \, t)}{\partial x_i \, \partial x_j} = e_{ij}(\overline{x}, \, t) \tag{17}$$

$$\frac{\partial A(\overline{x})}{\partial x_i} = a_i(\overline{x}) \tag{18}$$

$$\frac{\partial^2 A(\overline{x})}{\partial x_i \, \partial x_j} = b_{ij}(\overline{x}) \tag{19}$$

Then,

$$A(x) = A(\overline{x}) + \sum_{i=1}^{n} a_i(x_i - \overline{x}_i) + \tfrac{1}{2} \sum_{i,j=1}^{n} b_{ij}(x_i - \overline{x}_i)(x_j - \overline{x}_j) \tag{20}$$

$$u(x) = u(\overline{x}) + \sum_{i=1}^{n} v_i(x_i - \overline{x}_i) + \tfrac{1}{2} \sum_{i,j=1}^{n} z_{ij}(x_i - \overline{x}_i)(x_j - \overline{x}_j) \tag{21}$$

$$\text{where} \qquad u(\overline{x}) = \sum_{\mu, \nu=1}^{N} \eta_{\mu\nu} f(\overline{x}, t_{\mu}) S(t_{\nu}) \qquad (22)$$

$$v_i = \sum_{\mu, \nu=1}^{N} \eta_{\mu\nu} d_i(\overline{x}, t_{\mu}) S(t_{\nu}) \qquad (23)$$

$$z_{ij} = \sum_{\mu, \nu=1}^{N} \eta_{\mu\nu} e_{ij}(\overline{x}, t_{\mu}) S(t_{\nu}) \qquad (24)$$

or the limiting values of these as $\{t_{\mu}\}$ becomes dense in (T_1, T_2).

For the white noise case, the limiting values become

$$u(\overline{x}) = \frac{2}{N_o} \int_{T_1}^{T_2} f(\overline{x}, t) S(t) \, dt \qquad (25)$$

$$v_i = \frac{2}{N_o} \int_{T_1}^{T_2} d_i(\overline{x}, t) S(t) \, dt \qquad (26)$$

$$z_{ij} = \frac{2}{N_o} \int_{T_1}^{T_2} e_{ij}(\overline{x}, t) S(t) \, dt \qquad (27)$$

Now make the further assumption that

$$p(x) = (2\pi)^{-n/2} \Gamma^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \sum_{i,j=1}^{n} \gamma_{ij}(x_i - \overline{x}_i)(x_j - \overline{x}_j) \right\} \qquad (28)$$

where

$$\Gamma = \det | \gamma^{-1} |$$

In other words, $\overline{x}_i$ are now taken to be the a-priori means of x_i, and the a-priori joint distribution of the parameters is taken to be Gaussian.

We then have

$$\beta = \frac{\int_X \beta(x) \exp\left\{-\tfrac{1}{2}\sum_{i,j=1}^{n} \xi_{ij}(x_i - \bar{x}_i)(x_j - \bar{x}_j) + \sum_{i=1}^{n} \zeta_i(x_i - \bar{x}_i)\right\} dx}{\int_X \exp\left\{-\tfrac{1}{2}\sum_{i,j=1}^{n} \xi_{ij}(x_i - \bar{x}_i)(x_j - \bar{x}_j) + \sum_{i=1}^{n} \zeta_i(x_i - \bar{x}_i)\right\} dx} \tag{29}$$

where

$$\xi_{ij} = \gamma_{ij} + \tfrac{1}{2} b_{ij} - \tfrac{1}{2} z_{ij} \tag{30}$$

$$\zeta_i = -\tfrac{1}{2} a_i + v_i \tag{31}$$

In Eqs. (30) and (31), v_i and z_{ij} are linear functions of the observed data $S(t)$.

Now,

$$-\tfrac{1}{2}\sum z_{ij}(x_i - \bar{x}_i)(x_j - \bar{x}_j) + \sum \zeta_i(x_i - \bar{x}_i) \tag{32}$$

$$= -\tfrac{1}{2}\sum \xi_{ij}(x_i - \tilde{x}_i)(x_j - \tilde{x}_j) + \tfrac{1}{2}\sum \zeta_{ij} \tilde{x}_i \tilde{x}_j$$

where

$$\tilde{x}_i = \sum_{j=1}^{n} (\xi^{-1})_{ij} \zeta_j + \bar{x}_i \tag{33}$$

Thus

$$\hat{\beta} = \frac{\int_X \beta(x) \exp\left\{-\tfrac{1}{2}\sum_{i,j=1}^{n} \xi_{ij}(x_i - \tilde{x}_i)(x_j - \tilde{x}_j)\right\} dx}{\int_X \exp\left\{-\tfrac{1}{2}\sum_{i,j=1}^{n} \xi_{ij}(x_i - \tilde{x}_i)(x_j - \tilde{x}_j)\right\} dx} \tag{34}$$

Now let us further suppose that (ξ_{ij}) is a positive matrix. Then we can write

$$\hat{\beta} = E[\beta(x); \xi; \tilde{x}] \tag{35}$$

where $E[\beta(x); \xi; \tilde{x}]$ is the expected value of $\beta(x)$ with respect to a joint Gaussian distribution of $(x_1, \ldots, x_n)$ having means $\tilde{x}_i$ and inverse covariance matrix (ξ_{ij}).

III. VARIOUS SPECIAL CASES

First, suppose that $\beta(x)$ is a function of only one of the components, say

$$\beta(x) = g(x_i) \tag{36}$$

Then, clearly,

$$\hat{\beta} = \hat{g} = \frac{1}{2\pi \left[(\xi^{-1})_{ii} \right]^{\frac{1}{2}}} \int_{-\infty}^{\infty} g(x_i) \exp \left\{ -\frac{1}{2} \frac{(x_i - \tilde{x}_i)^2}{(\xi^{-1})_{ii}} \right\} dx_i \tag{37}$$

In particular, setting $\beta(x) = x_i$, then

$$\hat{x}_i = \tilde{x}_i \tag{38}$$

If we set $\beta(x) = x_i^2$, then

$$\widehat{x_i^2} = (\tilde{x}_i)^2 + (\xi^{-1})_{ii} \tag{39}$$

Similarly, the MMSE estimate of any power of x_i can be expressed as a function of $\tilde{x}_i$ and $(\xi^{-1})_{ii}$.

Next, suppose that we set

$$\beta(x) = x_i x_j \tag{40}$$

for any particular i and j.

Then

$$\widehat{x_i x_j} = \tilde{x}_i \tilde{x}_j + (\xi^{-1})_{ij} \tag{41}$$

By well-known formulas, the MMSE estimate of any function of the form

$$\beta(x) = \prod_{i=1}^{n} (x_i)^{k_i} \tag{42}$$

can be determined as a function of $\widetilde{x}_i$, $i = 1, \ldots, n$, and $(\xi^{-1})_{ij}$, $i,j = 1, \ldots, n$.

In the foregoing formulas, $\widetilde{x}_i$ is given by Eq. (33).

IV.   ADDITIONAL REMARKS ON THE COMPUTATIONAL IMPLICATIONS

It has previously been mentioned that the procedure which has been described is not necessarily limited in applicability to cases where the statistical conditions are satisfied such that the resulting estimates are really MMSE, any more than least squares procedures are limited to cases where they are really maximum likelihood.

Within reasonable limits, one may regard the procedure defined by the foregoing equations as generally applicable, even if the noise $\{\mathcal{E}(t)\}$ and the a-priori parameter distribution are not really Guassian. Moreover, one can use the white noise formulas for v_i, z_{ij}, etc., even if the errors $\{\mathcal{E}(t)\}$ are really correlated. Within reasonable limits, one can regard the matrices $\eta_{\mu\nu}$ and γ_{ij} as arbitrary choices of the designer of the computation procedure. Naturally, the statistical quality of the resulting estimates will be best if these represent the "true" inverse covariance matrices of the noise errors and the a-priori estimates, respectively; but it is in many cases desirable to trade statistical quality for computational convenience.

The procedure stated above can also be used as a step in an iterative procedure, in the following way. We can regard the quanti-ties $\bar{x}_i$ as the initial estimates in the iterative procedure.

Then, we can obtain $\hat{x}_i$ by means of Eqs. (38) and (33). Then, $\hat{x}_i$ can be used in place of $\bar{x}_i$ in precisely the same procedure--i.e., with v_i, z_{ij}, a_i, b_{ij} evaluated at $\hat{x}_i$ instead of $\bar{x}_i$. This would produce another estimate (with the same observational data) which might be denoted $(\hat{x}_i)'$. The procedure would then continue until the estimate settled.

REFERENCES

1. Swerling, P., "A Class of Signal Processing Procedures Suggested
    by Statistically Optimum Procedures," The RAND Corporation
    paper P-3029, November 1964.

2. Swerling, P., "Parameter Estimation Accuracy Formulas," IEEE
    Trans. Info. Theory, Vol. IT-10, No. 4, October 1964.

3. Swerling, P., "Parameter Estimation for Waveforms in Additive
    Gaussian Noise," Journ. Soc. Indust. Appl. Math. Vol. 7, No. 2,
    June 1959, pp. 152-166

4. Swerling, P., "Optimum Linear Estimation for Random Processes
    as the Limit of Estimates Based on Sampled Data," 1IR Wescon
    Convention Record, Part 4, 1958, pp. 158-163.